



# Pràctica 8.2: Web Scraping (XPath)

## Lliuraments

Els resultats d'aquesta part de la pràctica s'hauran d'entregar en format PDF i l'entrega pot ser a través de GIT\* o el moodle.

\* S'ha d'entregar l'enllaç del GIT al moodle.

## Guió

Amb l'ajuda de l'inspector d'elements del navegador, investiga com està formatada la pàgina <https://scrapepark.org/>. Aquesta pàgina està preparada per fer *web scraping*, de manera que les rutes per arribar als diferents elements no són trivials. Descobreix la ruta XPath per arribar a cada un dels elements que es demana tenint en compte només la informació que es proporciona a l'enunciat.

## Exercici 1

Per començar, clona el repositori de GIT que es troba en aquesta ubicació i executa el codi Python per veure quin resultat dona.

[https://github.com/pauitc/practica8\\_2](https://github.com/pauitc/practica8_2)

## Exercici 2

- a. Executa les següents rutes XPath i observa el resultat que dona cada una. A continuació, explica les diferències que hi ha entre cada resultat i raona per què produeixen resultats diferents.

- i. `node()` vs `text()`

Ruta 1: `//div[@class='attribution']/p/node()`

Ruta 2: `//div[@class='attribution']/p/text()`

La ruta 1 muestra todo lo que esté dentro de `<p>`, ya sea texto o `node`, en cambio en la ruta 2 sólo muestra el texto que esté dentro de `<p>`.

ii. Barra simple vs barra doble

Ruta 1: `//ul[@class='navbar-nav']/li/a/text()`

Ruta 2: `//ul[@class='navbar-nav']//li/a/text()`

La ruta 1 muestra los textos de enlaces que sean hijos directos de listas, en cambio en la ruta 2 muestra todos los textos de enlace ya sean hijos o no de listas.

- b. Representa, en forma d'arbre l'estructura XML que resulta d'avaluar la següent ruta XPath (pots ignorar els salts de línia i espais).

i. `(//div/h5) [6]`

```
<h5>
  <span>New Skateboard</span>
</h5>
```

ii. `//div[@class='carousel-item'] [1]//h1`

```
<h1>
  <span>
    <span>Discounts</span>
    <br>20% off</br>
  </span>
  <span id="all-products">On all our products!</span>
</h1>
```

### Exercici 3

- c. Descobreix la ruta que arriba al **correu** de contacte que es troba al **<footer>** de la pàgina. **Comença la ruta a l'etiqueta <html>**

`/html`

sales@mail.com

```
/html/body/footer/div[@class='container']/div[@class='row']/div[@class='col-md-4']/div[@class='full']/div[@class='fullinformatio-f']/p/span/text()
```

- d. Troba la ruta que arriba a l'**atribut src** de la següent imatge (n'hi ha una al **<footer>**, i una al **<header>**, pots escollir):



images/logo.svg

```
/html/body/div[@class='hero-area']/div[@class=header-section]/div[@class='container']/nav[@class='navbar.navbar-expand-lg.custom-nav-container']/a[@class='navbar-brand']/img/text()
```

- e. Troba la ruta fins a l'atribut **src** de les imatges amb **alt="Client"**.

```
images/client-one.png  
images/client-two.png  
images/client-three.png  
//img[@alt='customer']/
```

- f. Troba la ruta fins a l'**adreça** de la pàgina web **"Fake Street 123"**. Fes que l'adreça XPath parteixi la següent ubicació:

```
//div[@class='information-f']/p[1]/strong/text()
```

Fake Street 123

```
//div[@class='information-f']/p[1]/span/text()
```

- g. Troba la ruta que arriba fins al **<h5>** del **"New Skateboard 12"**. **[Pista:** busca la utilitat de la funció *normalize-space()* ].

```
<h5>                                <span>New Skateboard</span> 12  
</h5>
```

```
//div[@class='detail-box']/h5[normalize-space()='New Skateboard 12']
```

- h. Partint de la ruta de l'apartat anterior, Troba la ruta que arriba fins al **preu** (text) del **"New Skateboard 12"**.

12

```
//div[@class='detail-box']/h6/text()
```

## Exercici 4

Canvia la ruta a <https://scrapepark.org/table.html> . Amb l'ajuda del navegador, comprova què hi ha dins d'aquesta pàgina i troba la ruta XPath dels següents elements.

- i. Troba la ruta XPath a tots els **preus** dels **elements de color 'Blue'**. El resultat ha de ser el següent:

Blue

\$64

\$70

\$80

\$85

```
//tbody/tr[td[contains(text(), 'Blue')]]/td[@class='price']
```

- j. Troba la ruta que imprimeix **els preus del longboard** que es troben a la 4a columna de la taula **pintats en vermell**.

Longboard

\$80

\$85

\$90

\$62

\$150

```
//tbody/tr/td[position()=4 and contains(@style, 'color:red')]
```

- k. Indica el nom i color de l'article que **val \$110**. Comença l'expressió de la següent manera: **[pista]**: hauràs de fer servir l'operador “[” ]

```
//td[text()=' $110 ']
```

Skate

Special

```
//tbody/tr[td[text()=' $110 ']]/td[1]/text()
```

```
//tbody/tr[td[text()=' $110 ']]/td[2]/text()
```

- l. Troba la ruta a **tots els preus** dels objectes “Purple” **excepte el preu** que està pintat en vermell.

<td>Purple</td>

<td class="text-center">\$55</td>

```
<td class="text-center">$60</td>  
<td class="text-center">$72</td>  
//td[contains(text(),  
'Purple')]/following-sibling::td[not(contains(@style, 'red'))]
```